Deep learning-driven insights into super protein complexes for outer membrane protein biogenesis in bacteria

Mu Gao\textsuperscript{1*}, Davi Nakajima An\textsuperscript{2}, and Jeffrey Skolnick\textsuperscript{1*}

\textsuperscript{1}Center for the Study of Systems Biology, School of Biological Sciences

\textsuperscript{2}School of Computer Science, Georgia Institute of Technology, Atlanta, Georgia, USA

*Correspondence: mu.gao@gatech.edu (MG) and skolnick@gatech.edu (JS)
Abstract

To reach their final destinations, outer membrane proteins (OMPs) of gram-negative bacteria undertake an eventful journey beginning in the cytosol. Multiple molecular machines, chaperones, proteases, and other enzymes facilitate the translocation and assembly of OMPs. These helpers usually associate, often transiently, forming large protein assemblies. They are not well-understood due to experimental challenges in capturing and characterizing protein-protein interactions (PPIs), especially transient ones. Using AF2Complex, we introduce a high-throughput, deep learning pipeline to identify PPIs within the E. coli cell envelope and apply it to several proteins from an OMP biogenesis pathway. Among the top confident hits obtained from screening ~1,500 envelope proteins, we find not only expected interactions but also unexpected ones with profound implications. Subsequently, we predict atomic structures for these protein complexes. These structures, typically of high confidence, explain experimental observations, and lead to mechanistic hypotheses for how a chaperone assists a nascent, precursor OMP emerging from a translocon, how another chaperone prevents it from aggregating and docks to an β-barrel assembly port, and how a protease performs quality control. This work presents a general strategy for investigating biological pathways by using structural insights gained from deep learning-based predictions.
Introduction

A structural component unique to gram-negative bacteria is the outer membrane (OM), composed of lipopolysaccharides and phospholipids in an asymmetric bilayer with embedded lipoproteins and transmembrane β-barrel proteins [1]. The latter group, outer membrane proteins (OMPs), play vital functional roles, e.g., exchanging small molecules with the environment through their transmembrane β-barrel porins. OMPs are synthesized by cytosolic ribosomes, translocated across the inner membrane (IM), and finally delivered to the OM via the OMP biogenesis pathway [2].

The translocation and folding of OMPs involve many proteins, including essential and auxiliary ones that form multiple complexes in cooperation [3, 4]. Two core complexes are the SecYEG translocon [5, 6] and the β-barrel assembly machine (BAM) [7, 8]. SecYEG, a hetero-trimer comprised of the secretion channel SecY and two additional subunits SecE and SecG, is anchored to the inner membrane (IM) and is responsible for moving most cell envelope proteins across the IM [5]. Periplasmic and OM proteins enter the channel in SecY via the SecA-dependent translocation pathway [6]. They possess signal peptides recognized by SecA, which inserts a protein substrate into SecY and powers it through the channel using energy from ATP hydrolysis [9]. The signal peptide, located at the N-terminus of a substrate, contains a hydrophobic segment that folds into a transmembrane α-helix once it exits the lateral gate of SecY. To release the substrate from the IM, a type I signal peptidase (SPase I) cleaves the signal peptide [10]. At this point, a periplasmic protein has reached its destination, but an OMP, escorted by the chaperones SurA or Skp, continues its journey toward BAM harbored in the OM[11]. Five subunits constitute BAM, of which BamA plays the major role in folding a β-barrel and releasing the matured product [7, 8].

There are many open questions concerning the OMP biogenesis pathway. In E. coli, the SecYEG translocon recruits additional proteins, such as SecA [12], SecDF [13], YidC [14], PpiD [15, 16], to form a variety of supercomplexes in different scenarios. We do not know the identities of all members of these supercomplexes, let alone their atomic structures. SurA plays a major role in chaperoning a nascent OMP [11, 17]. How does it handle and deliver a substrate to BAM? Likewise, BAM recruits additional helpers, e.g., BepA [18], but how do they work together?

To answer these questions experimentally is challenging [19-22]. Recently, deep learning approaches have made tremendous progress in predicting the structures of protein complexes [23-26]. Here, we use one such a method to address the above questions. The centerpiece of our approach is AF2Complex [23], built on AlphaFold2 (AF2) [25, 27]. Using AF2Complex, we combine virtual screening for PPIs and supercomplex modeling and apply this strategy to several important proteins in the OMP biogenesis pathway.

Results

Virtual screening for protein-protein interactions in the E. coli envelopome

A workflow employing AF2Complex was implemented to search for interacting proteins within the E. coli cell envelope. We refer to these proteins collectively as the “envelopome”, which consists of ~1,450 proteins, or ~35% of the complete E. coli proteome. Given a query envelope protein sequence, screening is conducted against every envelope protein, including the query itself to test if it forms a homo-oligomer. For each pair of input proteins, AF2Complex predicts their structures simultaneously as a putative dimer and evaluates the interface score (iScore) of the resulting structural models to decide if they interact. The value of iScore ranges between 0 and 1. Interacting protein pairs typically return a positive iScore, whereas non-interacting proteins usually
return 0 or low iScores. Typically, 20 final structures are produced in multiple independent runs using ten different AF2 deep learning models. The highest iScore among all predicted structures of a target is used for ranking the entire envelopome.

According to previous benchmarks on a set of ~7,000 putatively non-interacting protein pairs from E. coli, minimum iScore thresholds of 0.40, 0.50, 0.70 yield false positive rates of 1.2%, 0.4%, and < 0.01%, respectively [23]. With respect to the quality of predicted structures, on a set of 440 heterodimeric complexes whose experimental structures have been recently determined and were not used for AF2 model training, the same three iScore thresholds predict medium- or high-quality complex structures for 84%, 87%, and 93% of the dimeric targets, respectively [23]. Hence, we consider predictions of medium, high, and very high confidence, progressively, at iScore cutoffs of 0.4, 0.5, 0.7. Given a query protein, the top iScores of all scanned envelope proteins are then sorted to identify potential interacting partners with the query protein. In this study, we select those with known biological relevance from high confidence predictions, and then predict the structures of larger complexes if more than two proteins are involved according to our predictions or based on literature information.

As a first test, we applied our PPI pipeline to the chaperones PpiD and YfgM, because strong experimental evidence indicates that these two proteins belong to a super SecYEG translocon [15, 16, 20, 28-32], yet their structures remain elusive. Fig. 1a shows the screening results of PpiD and YfgM against the envelopome. Reassuringly, these two proteins stand out at the very top of the predicted PPI rankings among the E. coli envelopome with iScores of ~0.73 in each case. Furthermore, PpiD hits SecY at the 12th rank (iScore = 0.43), corroborating previous experimental studies that SecY is an interacting partner of PpiD [15, 16, 29].

**SecYEG associates with the chaperone PpiD and assistant YfgM**

Subsequently, we modeled SecYEG together with PpiD and YfgM and acquired another very high confidence structure (iScore = 0.73, mean pLDDT = 79, see Methods) (Fig. 1b-e). Between this supercomplex and the dimer structures obtained from screening, the PPI interfaces between PpiD and YfgM are very similar according to iAlign[33], which yields a root-mean-square deviation (RMSD) of 0.6 Å for the CA atoms of the interface residues. Thus, we focus on the supercomplex structure below.

Individually, PpiD and YfgM in our complex model displays the same architecture as their respective monomer models calculated by applying AF2 [34]. PpiD by itself has four domains in an open-arm shape, fitting its proposed chaperone role [15, 31], plus an N-terminal transmembrane α-helix (Fig. 1b). The membrane anchoring helix is linked to a domain consisting of two discontinuous segments (residues 40–189, 581–623) from the terminal regions. Within this domain, two α-helices (residues 167–189, 581–609) are held by the tetratricopeptide repeat (TPR) domain of YfgM, which inserts itself into the IM via a single hydrophobic α-helix. Structurally, the PpiD/YfgM interface resembles a pair of “chopsticks” (two PpiD helices) grasped by a “palm” (the TPR domain). Their PPI interactions are quite extensive, including ~100 inter-protein residue-residue contacts involving both hydrophobic and polar partners (Fig. 1b,c), which may explain the strong binding affinity observed in experiments [20].

In addition to the channel that connects the cytosol to the periplasm, SecY features a lateral gate embedded in the IM. It opens to release transmembrane segments, including the signal peptide of a substrate such as OmpA [12]. *In vivo* photo cross-linking studies found numerous sites at the lateral gate probed by PpiD [16, 30]. Because PpiD has only a single transmembrane α-helix, a
logical speculation is that the α-helix interacts with the lateral gate of SecY. Indeed, our structure reveals that PpiD guards the lateral gate of SecY with its α-helix (Fig. 1b,d), which occupies the same position as a signal peptide leaving the SecY gate (Fig. S1). In superposition with a SecY crystal structure co-crystallized with an OmpA signal peptide, the Cα RMSD of SecY is 2.4 Å between the X-ray and computed structures. Thus, SecY in our model is open to accommodate the α-helix of PpiD, in comparison to an EM structure of E. coli SecY in which the gate is closed (Fig.
Our model further elucidates why SecY residues located in transmembrane helices (TM2/3/7/8) can be cross-linked to PpiD, because they are located at the PpiD/SecY interface. For example, strong cross-linking signals have been found for five SecY residues [16, 30], Ile91$_{\text{SecY}}$, Leu94$_{\text{SecY}}$, Leu127$_{\text{SecY}}$, Phe130$_{\text{SecY}}$, Phe286$_{\text{SecY}}$, and all make direct physical contacts (< 4.5 Å) with PpiD in our structural model (Fig. 1d). Moreover, we identified direct contacts between a hydrophobic pocket of PpiD and SecG, explaining the experimentally observed crosslinking of Phe122$_{\text{PpiD}}$ to SecG [29] (Fig. 1e). The interaction assigns SecG a functional role in coordinating with PpiD. Overall, the predicted structure rationalizes the results of several experimental studies.

### The chaperone PpiD interacts with the disulfide isomerase DsbA

Among the top hits for PPI partners of PpiD, a fascinating discovery is DsbA, ranked third with a high confidence iScore of 0.55. DsbA donates its disulfide bond to a substrate in need; thus, it is critical to the folding of a nascent protein as it leaves the SecY channel [35, 36]. How does DsbA coordinate with the SecY translocon? The predicted interaction immediately provides an answer, i.e., by associating with the chaperone PpiD that is part of a super translocon complex (SecYEG/PpiD/YfgM/DsbA). In a predicted structure of this supercomplex (iScore = 0.71, mean pLDDT = 78, Fig. 2a), PpiD uses its “hand” to grip DsbA, as observed in the top model of the pair, because their interactions do not interfere with other members of the translocon. The hand of PpiD is a parvulin-like domain, but it is devoid of the peptidyl-prolyl isomerase (PPIase) activity canonical to parvulin, due to the mutation of a critical histidine [37]. As our structure reveals, the catalytically inactive pocket of the PPIase domain is hydrophobic and buries the aromatic sidechain of Phe148$_{\text{DsbA}}$, which becomes a hot-spot residue surrounded by six hydrophobic residues of PpiD upon complexation (Fig. 2b). In addition, three pairs of salt-bridges are formed around the hydrophobic contacts, contributing to specific recognition. Notably, the catalytic cysteines of DsbA, Cys49$_{\text{DsbA}}$ and Cys52$_{\text{DsbA}}$, are on the opposite side to the protein-protein interface (Fig. 2a); thus, they are open to engage a substrate as it emerges from the SecY channel. Based on this model, we hypothesize that DsbA resides on PpiD transiently to improve its chance of encountering a substrate.

![Fig. 2. Structural model of the SecYEG/PpiD/YfgM/DsbA supercomplex. (a) Two views of the predicted structure. DsbA is shown in red, while the other proteins are colored the same as in Fig. 1. Two cysteines, Cys49 and Cys52, essential to the enzymatic function of DsbA, are shown as spheres. (b) PPI sites between PpiD and DsbA. For clarity, tertiary structures are transparent. Key interacting residues are shown in the licorice representation for PpiD and in the ball-and-stick representation for DsbA.](https://doi.org/10.1101/2022.08.25.505253)
SPase I LepB accesses an OMP substrate received by PpiD/YfgM

Polypeptide translocation triggers the dissociation of PpiD from SecY [16]. The modeled complex suggests that PpiD dissociation is realized by pushing a substrate out of the SecY gate, e.g., the signal peptide, which then repels the PpiD helix bound to the gate. For an OMP, a key next step is the removal of the signal peptide by the peptidase LepB in *E. coli* [10]. How does it operate? The dissociation of PpiD from SecYEG vacates a space for LepB to approach a substrate. We generated a structural model of LepB in complex with a precursor OmpA (proOmpA) chain (residue 1−87) grasped by PpiD (iScore = 0.64, mean pLDDT = 80) (Fig. 3). The OmpA cleavage site, Ala21OmpA, directly contacts the LepB active site triad, Ser89LepB, Ser91LepB, and Lys146LepB at ~4 Å away from the alanine. In our model, the proOmpA chain contributes a β-strand to form a parallel β-sheet with LepB, as has been speculated [10]. Overall, the structure provides a model for how LepB accesses a substrate received by PpiD/YfgM to cleave a signal peptide.

![Fig. 3. Predicted structure of the PpiD/YfgM/LepB/OmpA supercomplex.](image)

The chaperone SurA opens to load an OMP substrate

After passage through a SecYEG translocon, a nascent OMP polypeptide is relayed to SurA, which conveys the substrate toward its next stop, BAM. A crystal structure of SurA exhibits three domains: a core domain (split into N- and C-terminal subdomains), and two PPIase domains, P1 and P2 [38]. P2 retains PPIase activity but P1 does not [39]. In this crystal structure, P1 further packs with the core domain forming a cradle-like structure [38], referred to as the closed conformation of SurA. By contrast, another crystal structure of the SurA homodimer lacking P2s displays an open conformation, where two P1s are uncoupled from the core domains to hold a dodecapeptide at the dimeric interface between P1s [40]. To explore conformations of SurA, we predicted structures for a single, full-length SurA and obtain two major conformations (Fig. 4a). One conformation closely resembles the cradle-like closed conformation with a backbone Cα RMSD of 2.4 Å from the crystal structure. The other mimics the open conformation but incorporates P2 absent in the X-ray structure. Overall, the open conformation resembles a “three-
prong hook”, where both P1 and P2 swing away from their resting position in the closed conformation.

SurA functions mainly as a monomer but may act as a dimer for a large client [41, 42]. How does SurA chaperone a substrate such as OmpA? OmpA consists of a β-barrel (N-terminal) [43] and a periplasmic (C-terminal) domain [44]. To model the OmpA polypeptide, we reduced the size of multiple sequence alignments provided to AF2Complex and removed all its structural templates. Consequently, we generated a monomeric OmpA structure with a non-native, collapsed N-terminal domain and a native-like periplasmic domain (Fig. S2). In the presence of SurA, the periplasmic domain maintains the same fold, but remarkably, the non-native β-barrel region completely unravels and wraps around SurA (Fig. 4b–c). The SurA/OmpA dimer is not stable as different wrapping configurations were observed in two separate structures, leading to very low iScores of ~0.05, because the iScore favors stable, specific interactions. Nevertheless, the SurA/OmpA models appear physical and provide a hypothetical basis for how the chaperone SurA could prevent
a polypeptide chain from aggregating and present an unfolded polypeptide to BAM for its final assembly. Intriguingly, in both predicted complex structures, the OmpA polypeptide passes through a SurA crevice at its core domain, where Tyr126\textsubscript{SurA} consistently attracts a substrate aromatic residue with π-π interactions. This may be one mechanism by which SurA recognizes OMPs, which are enriched with aromatic residues [40, 45]. Moreover, that the disordered β-barrel region has few interactions with SurA could explain how the system avoids high energy expenditure for β-barrel release.

To identify other interacting partners, we searched the \textit{E. coli} envelopome for PPIs with SurA. Interestingly, SurA recognizes itself at a high confidence iScore of 0.53 (Fig. 4d). The top model of a SurA homodimer exhibits twofold rotational symmetry, in which P1 domains are swapped, leading to an open conformation quite different from the above one (Fig. 4e-f). This conformation, like a four-prong hook, provides SurA with another configuration to handle clients, potentially large ones, consistent with experimental conclusions [41, 42].

**SurA and BamA specifically recognize each other**

Excitingly, envelopome screening also confidently detected that SurA interacts with BamA, the anchoring subunit of BAM at the final stop of OMPs. To understand how SurA interacts with BAM, we subsequently used AF2Complex to probe potential interactions of SurA with all five BAM constituents (BamABCDE) and acquired a high confidence complex structure (iScore = 0.75, mean pLDDT = 84) (Fig. 5a-b). Because BAM has been extensively studied structurally [7, 46], we focus on describing its interaction with SurA. Overall, SurA mainly interacts with BamA,
with similar interactions observed in both the top supercomplex and the BamA/SurA dimer structures, in the N-terminal domains of both proteins (Fig. 5b-c). Gln23\textsubscript{SurA} to Val28\textsubscript{SurA}, largely missing in the crystal structures of SurA [38, 40], now form a β-sheet with the β2 strand of POTRA1, the N-terminal domain of BamA. In addition, hydrophobic contacts and two salt bridges are present in the same region. The structure explains the result that Asp26\textsubscript{SurA} is cross-linked to BamA [47], because Asp26\textsubscript{SurA} is located at the center of the SurA/BamA interface. Furthermore, SurA adopts a closed conformation in our model, but the P2 domain rotates ~45º from its crystal position [38] to engage additional contacts with BamA and BamE (Fig. 5d-e). For example, Asn336\textsubscript{SurA} forms a hydrogen bond with the backbone oxygen atom of Gly275\textsubscript{BamA}; in the same cross-linking study [47], Asn336\textsubscript{SurA} is implied to interact with BamA. These results invite a hypothesis that SurA and BamA initiate payload transfer via specific docking at their N-terminal domains.

**Metalloprotease BepA flips a lid in complex with BAM**

SurA led us to BamA, for which we conducted another envelopome PPI screening. The top six hits include all expected ones, the other four BAM factors and SurA, but unexpectedly, BepA

![Fig. 6. Structural models of BAM and BepA. (a) Computational PPI screening identifies BepA as a top hit to BamA. (b) Top structural model of the heterodimeric complex of BamA (green) and BepA (purple) in cartoon representation. The lid of BepA is colored red. The active sites with the protease domain of BepA are shown in a surface representation (orange). The five POTRA domains of BamA are labeled P1–P5. (c) Predicted structure of the BAM/BepA supercomplex. The lid of BepA extends to an open conformation. The image was created from the same viewpoint as b. (d) Specific residue-residue contacts between BamA and the TPR and protease domains of BepA. (e) Close-up views of the lid and the BamA β-barrel in the surface (top) and cartoon representations (bottom). A hydrophobic contact between Ala180\textsubscript{BepA} and Leu780\textsubscript{BamA} is shown as spheres, and the lateral gate of BamA is between the β1 and β16 strands (dark blue).](https://doi.org/10.1101/2022.08.25.505253)
ranked third with high confidence (Fig. 6a-b). BepA was not anticipated because we were not aware of its relevance. However, a literature search quickly revealed that BepA is highly relevant because it cleans up stalled OMP folding at BAM [18, 48] and its interactions with BAM have been documented [18, 49], though no structure of the complex is available. Consequently, we modeled BepA and the full BAM complex altogether, obtaining a high confidence model (iScore = 0.68, mean pLDDT = 85, Fig. 6c). Overall, the BepA/BamA interfaces are quite similar (interface Cα RMSD 1.2 Å) between the dimer and the supercomplex structure. Extensive contacts are present between all five periplasmic domains (POTRA1−5) of BamA and the two domains (protease and TPR) of BepA. They can be clustered into two main groups scattered between the POTRA1 and TPR, and between POTRA3−4 and the protease domains (Fig. 6d). Strong photo cross-linking signals to BamA have been observed previously for several residues of TPR including Phe404BepA [18, 49]. According to our model, Phe404BepA belongs to the first interaction cluster with five BamA residues, particularly Pro47BamA. Another residue cross-linked to BamA is Gln428BepA, which establishes a specific hydrogen bond with Arg49BamA in our model.

Crystal structures of BepA alone have been solved, but a segment (residue 154 to 192) of the metalloprotease domain is absent in these structures, presumably due to its flexibility [49-51] (Fig. S3). The segment, termed an “active-site lid”, was speculated to be important to the protease function of BepA, because its movement could either expose (open state) or cover (closed state) the catalytic site, the HEXXXH motif [50]. Intriguingly, we only see the closed conformation in 40 models of the BamA/BepA heterodimer (Fig. 6b), but both the open and closed states in 8 supercomplex models. In the open state, two α-helices comprising the lid are in a straight-up configuration (Fig. 6c), whose upper half (residue 157 to 186) rotates ~105° to cover the enzymatic site in the closed configuration (Fig. 6b). Moreover, the flexible lid, with a large hydrophobic surface similar to the β-barrel of BamA, is in a position as if it is inserted into the inner leaflet of the OM and is in the proximity of the lateral gate of BamA (Fig. 6e). As such, BepA could control its protease activity via probing a substrate β-barrel stemming out of BamA. Hence, it is hypothesized that a budding β-barrel blocks the lid opening, whereas a stalled OMP within BamA permits lid opening and subsequent substrate cleavage.

Discussion
Here we have demonstrated a deep learning strategy that combines virtual PPI screening over the E. coli envelopome and supercomplex structure modeling. By applying it to several key proteins in the OMP biogenesis pathway, we have identified their functional partners within the top 1% ranking of ~1,450 proteins screened for PPIs per query. Thanks to high confidence structures underlying the top predictions, one can understand many experimental phenomena, particularly in vivo site-directed photo cross-linking data. For example, cross-linked products found from the SecYEG or BAM supercomplexes may be explained by direct physical interactions revealed in our predicted structures. Moreover, previously speculated conformations are captured for SurA and BepA. Most importantly, these revealing atomic structures suggest mechanistic hypotheses for various steps of the OMP biogenesis pathway as summarized in Fig. 7, where we present their diagrams along with some predicted supercomplex structures.

One unexpected discovery is the DsbA/PpiD interaction. It was known that DsbA crucially transfers its disulfide bond to a nascent polypeptide translocated by SecYEG [35, 36], but how does the translocon interact with DsbA? The predicted supercomplex of SecYEG/PpiD/YfgM/DsbA provides a compelling answer for three reasons: First, the transient
residence of DsbA on PpiD at the translocon dramatically improves its efficiency versus a random search in a crowded periplasm. Second, it predicts a function for the PPIase domain of PpiD that mysteriously lacks the PPIase function. Third, since PpiD is anchored to the IM, it can keep DsbA close to DsbB, an IM protein that recycles DsbA. Notably, two other *E. coli* chaperones, SurA and FkpA, possess two PPIase domains each, and AF2Complex predicts that DsbA interacts with FkpA (iScore = 0.47 versus 0.55 for PpiD) but not with SurA (iScore = 0.03).

Interestingly, the AF2 neural networks can model a partially unfolded polypeptide accompanied by chaperones, even though AF2 was trained on folded proteins. We attained structural models that appear to be at least partially physical. One example is the model of
PpiD/YfgM/LepB/proOmpA, in which proOmpA is posed for cleavage by peptidase LepB. The cleavage Alanine of OmpA is ~4 Å away from the catalytic triad of LepB. Considering that only apo or inhibitor bound structures of LepB are available, this computed model implies that AF2 has learned physical representations. More intriguing examples are the models of SurA/OmpA, where the periplasmic domain of OmpA is folded but the β-barrel domain is completely unfolded and loosely wrapped around SurA. The predicted structures echo the NMR structures of an unfolded polypeptide bound to SecB, a cytosolic chaperone involved in the early stage of the SecA-dependent translocation pathway [52]. Despite the low confidence due to weak interactions, the predicted structures delineate a picture for how SurA prevents OmpA from aggregating. Moreover, since it transports OmpA with a relatively small number of intermolecular contacts, the free energy required to dissociate OmpA from SurA is small.

These results reinforce the notion that deep learning is a promising way to explore the conformational ensemble of proteins and to uncover the molecular mechanisms of biosystems [23]. The combination of advanced AF2 deep learning models, an effective PPI ranking metric, and a workflow optimized for large-scale screening generates illuminating structures of protein complexes. In the presented examples, we focused on those with obvious biological relevance according to the literature. There are of course other confident PPI predictions that were not described here. While some are promiscuous interactions or physically possible but biologically irrelevant, or simply false, there are likely functional interactions yet to be explored. More generally, these results are an example of a deep learning-based strategy that help elucidate mechanistic aspects of complex biochemical pathways.
Methods

**E. coli cell envelopome.** The proteome for *E. coli* strain K12 MG1655, consisting of 4,400 protein sequences was retrieved from UniProt [53] in March 2022 (Proteome ID UP000000625). Then, the subcellular location of each protein was parsed to collect all known and predicted cell envelope proteins. To be conservative, we included all proteins whose primary subcellular location is not the cytosol, yielding 1,466 proteins defined as the envelopome. The set contains all proteins located in the inner membrane, periplasm, outer membrane, and extracellular surface, and some cytosolic proteins located primarily at the periphery of the periplasmic inner membrane, e.g., SecA. For AF2Complex modeling, we removed 11 sequences that contain at least one non-standard amino acid.

**AF2Complex.** An updated version of AF2Complex was built upon AlphaFold version 2.2.0[25]. AF2Complex supports three different set of deep learning neural network models provided with AF2, i.e., the original models for monomer prediction (AF version 2.0.1 [27]) and two set of models for multimer prediction (version 2.1.0, and version 2.2.0 [25]). Numerous changes were made; we list three major ones here. First, the interface score (iScore) metric was introduced to rank the confidence of a predicted complex model [23]. Second, the data pipeline for generating input features was split from the neural network inference. This allows rapid assembly of input features for individual proteins of a putative complex from pre-generated input features of the full proteome. Third, five options for multiple sequence alignment (MSA) pairing are provided: no pairing, all paired, cyclic, linear, and arbitrary pairing. The last three options are experimental, and we only employed the first two MSA pairing modes in this work. All structural models, either of a monomer or a multimer, were predicted by AF2Complex using either the ‘monomer_ptm’ deep learning models without any MSA pairing or the ‘multimer_v2’ models with all MSA pairing, the latter only used for structure prediction of complexes. The input features of all monomers were derived using standard sequence libraries [54-56] and a version of the Protein Data Bank[57] released in November 2021. The confidence of an output structure is evaluated using three different metrics: the interface score for protein interface [23], the predicted TM-score for global structure of a monomer [58], and the predicted local distance difference test (pLDDT) score for local domain structure [27, 59]. According to AF2, a mean pLDDT score higher than 70 indicates high confidence in a predicted structure when evaluated on individual domain(s).

**Envelopome PPI screening.** We adapted a workflow for proteome-scale monomeric structure prediction with AF2 implemented on the Summit supercomputer at Oak Ridge National Laboratory [60]. The workflow was applied to four query proteins: PpiD, YfgM, SurA, and BamA. Given a query sequence and an envelopome protein sequence, AF2Complex assembles the pre-generated monomeric input features and then feeds the composite features to the AF2 neural network models for inference. Ten different neural network models are used as mentioned above, and each deep learning model is invoked twice with two different random seeds and up to eight recycles. This procedure typically gives 20 final structural models per target for ranking, and the top ranked model by iScore is retained. Although the ‘multimer_v2’ set of models greatly reduces unphysical clashes in predicted structures compared to structures predicted by the original set of multimer deep learning models, there is still a small chance of generating severe clashes, which could yield an artifactually high iScore. We filter out these structures by applying a minimum interface clash indicator of 0.4 [23]. We did not apply model relaxation to these predicted structures because most are not actual protein complexes; as such, it is not worth expending the extra computing time. To avoid memory overflow, we limited the input MSA depth to 5,000 for
each monomer and allowed a maximum of four structural templates per monomer. A maximum of 1,600 amino acids was imposed on the total size of each complex target. This limitation slightly reduced the total number proteins in the envelopome for screening, typically to ~1,450 for the four proteins studied. Because about 30% of envelopome proteins have a signal peptide that is absent in their mature chains, only mature chains were used for screening. This is readily realized with AF2Complex because it provides an option to crop arbitrary segments of a monomer input feature during feature assembly. The residue ranges of the mature chains were obtained from the UniProt knowledgebase.

**Modeling the OmpA polypeptide.** It was necessary to model OmpA as a substrate of PpiD or SurA. OmpA was chosen mainly because it is a model OMP for studying the OMP biogenesis pathway [61]. Many experimental data exist for validation. To minimize potential “memory” effects due to large MSAs or structural templates, we reduced the number of sequences in the MSAs and removed all structural templates in the input features of OmpA. We tested MSA depths of 1, 10, 20, 50, and 100 to predict a structural model of proOmpA in complex with SecYEG. The goal is to generate a structural model that mimics the crystal structure of SecYEG translocating a proOmpA polypeptide [12]. Our tests found that an MSA depth below 50 can yield a model structure of an unfolded OmpA peptide through the channel of SecY, when used with either ‘model_1_p tm’ or ‘model_3_p tm’ of all AF2 neural network models. In this study, we used an MSA depth of 20 and the two AF2 models to predict structural models involving OmpA.

**Computational resources.** *E. coli* envelopome PPI screening was performed on the Summit supercomputer, typically using 923 nodes for several hours of wall clock time. Each Summit node hosts 6 Nvidia 16 GB V100 GPUs. The structure predictions of various supercomplexes were conducted locally using about 10 workstations each with four Nvidia RTX6000 GPUs, each with 24 GB of GPU memory.

**Analysis.** The program VMD[62] was used to inspect predicted structural models and create all molecular images. APoc was used to align monomeric protein structures [63], and iAlign was used to perform protein-protein interface comparison [33].

**Data and Code availability**

The input features for the full *E. coli* proteome used for envelopome screening by AF2Complex, and the computational models presented in this study are available at Zenodo https://doi.org/10.5281/zenodo.6846915. The source code of AF2Complex is freely available at https://github.com/FreshAirTonight/af2complex.
References


51. Shahrizal, M., Y. Daimon, Y. Tanaka, Y. Hayashi, S. Nakayama, S. Iwaki, et al., 
Structural basis for the function of the β-barrel assembly-enhancing protease BepA. J. 
52. Huang, C., P. Rossi, T. Saio, and C.G. Kalodimos, Structural basis for the antifolding 
Universal Protein Resource (UniProt): an expanding universe of protein information. 
D570-D578.
p. 590-596.
58. Zhang, Y. and J. Skolnick, Scoring function for automated assessment of protein 
59. Mariani, V., M. Biasini, A. Barbato, and T. Schwede, lDDT: a local superposition-free 
score for comparing protein structures and models using distance difference tests. 
Proteome-scale Deployment of Protein Structure Prediction Workflows on the Summit 
61. Reusch, R.N., Insights into the structure and assembly of Escherichia coli outer 
Acknowledgments: We thank Jerry M. Parks for stimulating discussions and critical reading of the manuscript. This work was supported in part by the DOE Office of Science, Office of Biological and Environmental Research (DOE DE-SC0021303) and the Division of General Medical Sciences of the National Institute Health (NIH R35GM118039). The research used resources supported in part by the Advanced Scientific Computing Research (ASCR) Leadership Computing Challenge (ALCC) program, and by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology.

Author contributions: JS and MG conceived the project, DNA and MG implemented the source code, MG performed PPI screening and structural modeling, analyzed the data, and prepared the first draft of the manuscript, JS and MG revised the manuscript, and all authors proofread the manuscript.

Competing interests: Authors declare no competing interests.
**Fig. S1.** Comparison of a computed structure of SecY (silver) and two experimental structures (magenta).  
**a,** Superposition of an X-ray crystal structure of SecY from *G. thermodenitrificans* (PDB code: 5EUL) onto SecY from the predicted structure of the *E. coli* SecYEG/PpiD/YfgM supercomplex. In the experimental structure, SecY was co-crystalized with the signal peptide (green) of a precursor OmpA. In the computed structure, the N-terminal transmembrane α-helix (blue) of PpiD is also displayed. All structures are shown in the cartoon representation.  
**b,** Superposition of an EM structure of SecY from *E. coli* (PDB code: 5GAE, chain g) onto the computed model. For clarity, the α-helix of PpiD was omitted.
**Fig. S2.** Structural models of the OmpA polypeptide in the absence of SurA. **a,** Predicted model of OmpA obtained with a shallow MSAs and no structural templates. The N-terminal β-barrel domain is collapsed but not in its native fold; the C-terminal periplasmic domain is native-like. Two cysteines, Cys311 and Cys323, forming a disulfide bond are shown as spheres. **b,** The predicted model aligned to an NMR structure of the periplasmic domain (PDB code: 2MQE, TM-score [58] ~0.75).
Fig. S3. Predicted structures of BepA compared to two experimental structures. **a, b,** Superposition of structural models of the lid (red/magenta) in closed and open states, respectively, onto a crystal structure (PDB code: 6AIT, cyan). **c,** Superposition onto another crystal structure (PDB code: 6ASR, cyan).